C'est comme du cours aujourd'hui !

04/09/2026

Modèle linéaire

$\mathcal{D}_n = (X_i, Y_i)_{1 \le i \le n}$, $X_i \in \mathbb{R}^p$, Y_i : output réel $\in \mathbb{R}$.

On souhaite faire de la régression : trouver 1 fonction f telle que $Y \simeq f(X)$, f est calibrée à partir du jeu de données $\mathcal{D}_n$.

Stratégie 1 : utiliser géom + optim pour trouver f.

Ex Méthode des moindres carrés :

- on met 1 structure sur f, par exemple un mod. linéaire : $f_\beta(x) = \langle \beta, x \rangle$
- on définit un critère d'être "le meilleur" f pour approximer Y : $\sum_{i=1}^{n} (y_i - f_\beta(x_i))^2$ est minimal.

Stratégie 2 : On met un modèle statistique.

$Y = \langle X, \beta \rangle + \sigma \varepsilon$ avec β et σ des paramètres du modèle. On cherche alors à estimer β et σ via les obs. $\mathcal{D}_n$. $\varepsilon \sim \mathcal{N}(0,1)$, Y_i indépendantes

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{ p_\theta \cdot \mathrm{dLeb}^{\otimes n} := \bigotimes_{i=1}^{n} \underbrace{\mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)}_{\langle \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} 1 \\ x_i \end{pmatrix} \rangle + \sigma \varepsilon} : \theta := (\beta_1, \beta_2, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+^* \right\} \right).$$

$\downarrow$

Y_i sont $\perp\!\!\!\perp$.

Ici, on a $p = 1$ et $Y_i \simeq \beta_1 + \beta_2 x_i + \sigma \varepsilon_i$.

$$Y_i = \langle X_i, \beta \rangle + \sigma \varepsilon .$$

$X_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$

1. Montrer que la matrice $\mathbf{X^T X}$ est inversible.
(*Remarque : Puisque $u^T \mathbf{X^T X} u \geq 0$ pour tout $u \in \mathbb{R}^2$, on en déduit que la matrice est définie positive.*

↳ C'est vrai tout le temps, il n'y a pas besoin que $p = 1$

$X^T X$ est toujours ≥ 0

$\forall u$: $u^T X^T X u = \|Xu\|^2 \geq 0$. Donc $X^T X$ est ≥ 0

$X^T X$ est de taille 2×2.

$$X^T X = \begin{bmatrix} n & \langle 1, x \rangle \\ \langle 1, x \rangle & \|x\|^2 \end{bmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\det(X^T X) = n \sum x_i^2 - \left(\sum x_i\right)^2$$

≥ 0 par Cauchy-Schwarz et le cas d'égalité est lorsque les vecteurs sont colinéaires

On a $\det(\,) = 0 \iff 1$ et x st col.

$\iff$ tous les x_i sont égaux

Or, $\exists\, i \neq j : x_i \neq x_j$. Donc la conclusion.

Rem : $X^T X$ est une matrice de Gramm.
Cette matrice est inversible lorsque X est de rang maximal, p
(en admettant que $n > p$)

2. Déterminer les estimateurs du maximum de vraisemblance $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$ de $(\beta_1, \beta_2, \sigma^2)$.

On va étudier le contexte général : $Y_i = \langle X_i, \beta \rangle + \sigma \varepsilon_i$

$\varepsilon_i \sim \mathcal{N}(0, 1)$ $\qquad \theta = (\beta, \sigma)$

$$\theta \longmapsto L_n(\theta) = \prod_{i=1}^{n} f_\theta(Y_i \mid X_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(Y_i - \langle X_i, \beta\rangle)^2}{2\sigma^2}}$$

$$\ell_n(\theta) = \log L_n(\theta)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum (Y_i - \langle X_i, \beta\rangle)^2}{2\sigma^2}$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X\beta = \begin{pmatrix} \langle X_1, \beta\rangle \\ \vdots \\ \langle X_n, \beta\rangle \end{pmatrix} \quad \text{avec} \quad X = \begin{pmatrix} x_1^{(1)} \cdots x_1^{(p)} \\ \vdots \\ x_n^{(1)} \cdots x_n^{(p)} \end{pmatrix}$$

X : c'est la matrice de design du modèle linéaire.

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\|Y - X\beta\|^2}{2\sigma^2}$$

Supposons β connu : on observe $\sigma^2 \longmapsto \ell_n(\theta)$

$$\partial_{\sigma^2} \ell_n(\theta) = -\frac{n}{2\sigma^2} + \frac{\|Y - X\beta\|^2}{2\sigma^4}$$

$$\partial_{\sigma^2} \ell_n = 0 \iff \widehat{\sigma^2} = \frac{\|Y - X\beta\|^2}{n}$$

β sera celui qui maximise

Etudions l'équation en β :

$$\ell_n(\theta) = \text{Cte}(\sigma) - \frac{\|Y - X\beta\|^2}{2\sigma^2}$$

On doit minimiser $\|Y - X\beta\|^2 := J(\beta)$

$\beta \longmapsto J(\beta)$ est $\mathcal{C}^\infty(\mathbb{R}^p, \mathbb{R})$

On calcule $\nabla J(\beta)$ et on cherche le ou les points critiques.

On observe en + que comme $\beta \longmapsto X\beta$ parcourt $\text{Vect}(X)$ alors le min est atteint pour $X\beta = \underset{\text{Vect}(X)}{\text{Proj}}(Y)$.

$\nabla J(\beta)$ est l'unique vecteur d_β tq

$$J(\beta+h) = J(\beta) + \langle h, d_\beta \rangle + o(|h|)$$

$$J(\beta+h) = \|X(\beta+h) - Y\|^2$$

$$= \|X\beta - Y\|^2 + 2\langle Xh, X\beta - Y\rangle + \underbrace{\|Xh\|^2}_{o(|h|)}$$

$$= J(\beta) + 2\langle h, \underbrace{X^T(X\beta - Y)}_{\nabla J(\beta)}\rangle + o(|h|)$$

On en déduit que $\boxed{\nabla J(\beta) = 2\, X^T(X\beta - Y)}$

Pour la Hessienne : - soit on dérive le gradient

- soit on fait Taylor ordre 2.

$$D^2 J(\beta) = 2\, X^T X$$

On observe que si X^TX est sym def ≥ 0, alors J est convexe strictement.

On résout $\nabla J(\beta) = 0 \iff X^TX\beta - X^TY = 0$

$$\iff X^TX\beta = X^TY$$

Formule à connaître par coeur. $\iff \boxed{\hat{\beta} = (X^TX)^{-1}X^TY}$

Dans le cadre de l'exercice :

$$(X^TX)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

$$X^TY = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$(X^T X)^{-1} X^T Y = \frac{\begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{pmatrix}}{n \sum x_i^2 - (\sum x_i)^2}$$

Dans la logique de la régression, on peut toujours se ramener à ce nuage de points centrés : $\sum x_i = 0$

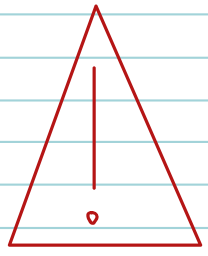
Dans ce cas là : $\hat{\beta}_1 = \frac{1}{n} \sum y_i$: moyenne des y

$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$: $\frac{\text{cov des } x \text{ avec } Y}{\text{Var}(x)}$

3. Déterminer la loi de $(\hat{\beta}_1, \hat{\beta}_2)$ sous $p_\theta \cdot \text{dLeb}^{\otimes n}$.

Pour déterminer la loi de $\hat{\beta}$, on va toujours utiliser

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$



Il y a un (β^*, σ^{*2}) qui sert à générer les données.
Je ne les connais pas, et je les cherche.
Il y a aussi $(\hat{\beta}, \hat{\sigma}^2)$ et je cherche leurs lois.

$$Y = X\beta^* + \sigma^* \underbrace{N(0, I_n)}_{\xi}$$

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta^* + \sigma^* \xi)$$

$$= \beta^* + \sigma^* (X^T X)^{-1} X^T \xi$$

$$\text{Cov}\left(\sigma^* (X^T X)^{-1} X^T \xi\right) = \sigma^{*2} \mathbb{E}\left((X^T X)^{-1} X^T \xi \cdot \xi^T X (X^T X)^{-1}\right)$$

$$\text{Cov}(A\xi) = \mathbb{E}(A \xi \xi^T A^T)$$

$$= \sigma^{*2} (X^T X)^{-1} X^T \underbrace{\mathbb{E}[\xi \xi^T]}_{I_n} X (X^T X)^{-1}$$

$$= \sigma^{*2} (X^T X)^{-1}$$

$$\hat\beta \sim \mathcal{N}\left(\beta^*, \sigma^{*2}(X^TX)^{-1}\right)$$

$\mathbb{E}[\hat\beta] = \beta^*$ $\hat\beta$ est sans biais.

$\mathbb{E}\|\hat\beta - \beta^*\|^2 = \sigma^{*2}\,\mathrm{Tr}\left((X^TX)^{-1}\right)$

4. Montrer que, sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$, $\hat\beta_1$ et $\hat\beta_2$ sont indépendants si et seulement si $n^{-1}\sum_{i=1}^n x_i = 0$

On sait que $\hat\beta$ est un vecteur Gaussien.

$\hat\beta_1 \perp\!\!\!\perp \hat\beta_2 \iff \mathrm{Cov}(\hat\beta_1, \hat\beta_2) = 0$

$\iff \sum x_i = 0$ (voir expression de $(X^TX)^{-1}$)

5. Déterminer la loi de $\hat\sigma^2$ sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$.

On sait d'après le 2°) que :

$$\hat\sigma^2 = \frac{\|Y - X\hat\beta\|^2}{n} = \frac{\|(X\beta^* + \sigma^*\xi) - X(X^TX)^{-1}X^T(X\beta^* + \sigma^*\xi)\|^2}{n}$$

On a remplacé $\hat\beta$ par sa valeur en fonction de β^*, σ^* et ξ idem avec Y.

$$\hat\sigma^2 = \frac{1}{n}\left\| X\beta^* + \sigma^*\xi - X(X^TX)^{-1}X^TX\beta^* - X(X^TX)^{-1}X^T\sigma^*\xi \right\|^2$$

$$= \frac{\sigma^{*2}}{n}\|\xi - X(X^TX)^{-1}X^T\xi\|^2$$

$$= \frac{\sigma^{*2}}{n}\|(I - P)\xi\|^2 \quad \text{où } P = X(X^TX)^{-1}X^T$$

P est une projection : $P^2 = P$; P est $\perp$: $P^T = P$ } sur $\mathrm{Vect}(X)$

$I - P$ est aussi la matrice de proj $\perp$ sur $\mathrm{Vect}(X)^\perp$.

$\|(I-P)\xi\|^2 \sim \chi^2(\mathrm{Rank}(I-P))$

$\mathrm{Rank}(P) = p$. $\mathrm{Rank}(I-P) = n-p$.

$$\mathcal{L}(\hat{\sigma}^2) = \frac{\sigma^{*2}}{n}\,\chi^2(n-p)$$

$$\mathbb{E}(\hat{\sigma}^2) = \frac{n-p}{n}\,\sigma^{*2}$$

L'estimateur sans biais associé est $\frac{1}{n-p}\,\|Y - X\vec{\beta}\|^2$

6. Montrer que sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$, $\hat{\sigma}^2$ et $(\hat{\beta}_1, \hat{\beta}_2)$ sont indépendants.

On le sait grâce à $(I-P)P = P(I-P) = 0$

Exercice 2 (Estimation d'un coefficient de mélange ($\star$)). Soit un n-échantillon $(X_1, \ldots, X_n)$ du modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{f_\theta \cdot \mathrm{Leb}, \theta \in]0,1[\})$ où

$$f_\theta(x) = \frac{\theta}{a}\,\mathbb{1}_{\{[0,a]\}}(x) + \frac{(1-\theta)}{b}\,\mathbb{1}_{\{[0,b]\}}(x).$$

On suppose que a et b sont connus et que $0 < a < b$. La fonction de répartition associée à la densité f_θ vaut

$$x \mapsto \begin{cases} 0 & x \le 0 \\ \left(\frac{\theta}{a} + \frac{(1-\theta)}{b}\right) x & 0 \le x \le a \\ \theta + \frac{(1-\theta)}{b} x & a \le x \le b \\ 1 & x \ge 1 \end{cases}$$

1. Exprimer la fonction de vraisemblance à l'aide de $N_a := \sum_{i=1}^n \mathbb{1}_{X_i \in [0,a]}$ le nombre d'observations à valeur dans $[0.a]$.

Les obs sont iid

$$L_n(\theta) = \prod_{i=1}^n f_\theta(X_i)$$

$$= \prod_{i=1}^n \left(\frac{\theta}{a}\,\mathbb{1}_{0 \le X_i \le a} + \frac{1-\theta}{b}\,\underbrace{\mathbb{1}_{0 \le X_i \le b}}_{=1} \right)$$

$$= \prod_{i=1}^n \left(\left(\frac{\theta}{a} + \frac{1-\theta}{b}\right)\mathbb{1}_{X_i \in [0;a]} + \frac{1-\theta}{b}\,\mathbb{1}_{X_i \notin [0;a]} \right)$$

$$= \left(\frac{\theta}{a} + \frac{1-\theta}{b}\right)^{N_a} \cdot \left(\frac{1-\theta}{b}\right)^{n-N_a}$$

$$\ell_n(\theta) = \log L_n(\theta) = N_a \log\left(\frac{\theta b + a(1-\theta)}{ab}\right) + (n-N_a)\log\frac{1-\theta}{b}$$

$\hat{\theta} = \arg\max \ell_n(\theta)$. On obtient alors :

$$\frac{N_a (b-a)}{\theta b + (1-\theta) a} - \frac{(n - N_a)}{1-\theta} = 0$$

$$N_a (b-a)(1-\theta) = (n - N_a)(\theta b + (1-\theta) a)$$

$$N_a (b-a) - (n-N_a) a = \theta \left((n-N_a) b - (n-N_a) a + N_a (b-a) \right)$$

$$N_a b - n \cdot a = \theta \, n (b-a)$$

$$\hat{\theta} = \frac{N_a b - n a}{n (b-a)} \vee 0$$

Oublions le max avec 0.

$$\mathbb{E}[\hat{\theta}] = \frac{b \, \mathbb{E} N_a - n a}{n (b-a)}$$

On observe que :

$$\mathbb{E}\left(\mathbb{1}_{X_i \in [0;a]} \right) = \theta^* + (1-\theta^*) \cdot \frac{a}{b}$$

D'où

$$\mathbb{E}[\hat{\theta}] = \frac{n b \left(\theta^* + (1-\theta^*) \frac{a}{b} \right) - n a}{n (b-a)}$$

$$= \frac{n b \theta^* + n a (1-\theta^*) - n a}{n (b-a)}$$

$$= \theta^*$$

3. En utilisant la méthode des moments (d'ordre 1), proposer un estimateur $\hat{\theta}_n^{(2)}$ de θ.

MOM method of moments

$$\mathbb{E}[X] = \theta \cdot \frac{a}{2} + (1-\theta) \frac{b}{2} = \frac{b}{2} + \theta \left(\frac{a-b}{2} \right)$$

$\frac{1}{n}\sum X_i \simeq \mathbb{E}X$ LFGN.

On résout $\frac{b}{2} + \theta\,\frac{a-b}{2} = \frac{\sum x_i}{n}$

$$\hat{\theta}_n^{(2)} = \left(\frac{2\sum x_i}{n} - b\right)\Big/(a-b)$$

$$= \left(b - \frac{2\sum x_i}{n}\right)\Big/ b-a$$